QUALITY ASSURANCE/QUALITY CONTROL FOR HIGH THROUGHPUT BIOASSAY PROCESS

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims priority under 35 U.S.C. § 119(e) to U.S. Provisional Patent Application No. 60/389,831, entitled "Quality Assurance/Quality Control for SELDI-TOF Mass Spectra," filed on July 29, 2002, the contents of which are hereby incorporated by reference in its entirety.

STATEMENT OF FEDERALLY SPONSORED RESEARCH

[0002] The research work described here was supported under a Cooperative Research and Development Agreement (CRADA) between the US Government and Correlogic Systems, Inc.

BACKGROUND OF THE INVENTION

[0003] The present invention relates generally to the field of bioinformatics. More specifically, the present invention relates to a method of quality assurance/quality control ("QA/QC") for bioinformatic systems.

[0004] Methods of analyzing biological samples are generally known. In a typical analysis, mass spectroscopy is performed on the biological sample to determine its overall biochemical make-up. Based on the mass spectra obtained from the mass spectroscopy, various diagnostics may be run.

[0005] When biological samples are analyzed, it is desirable to run more than one trial on the biological sample, thereby improving the accuracy of the diagnostic. Analysis of biological samples may be performed by using various high-throughput mass spectrometry related bioassay processes. A process can include using matrix assisted laser desorption ionization time-of-flight (MALDI-TOF) or electrospray techniques (i.e., generation of droplets by applying a high voltage to a stream of liquid). When performing multiple mass spectral analyses on the same sample, however, the spectra obtained can vary. This variation may be due to the mass spectrometer itself, from inconsistencies in the sample, heterogeneity in the patient population, or in sample handling and processing. A process that employed a protein chip or surface enhanced type of

mass spectrometry (SELDI-TOF) indicated that various chips yielded spectra that were inconsistent with one another. Similar effects were observed with respect to spectra obtained using electrospray techniques. This inconsistency can lead to inaccurate results when running a diagnostic.

SUMMARY OF THE INVENTION

[0006] The present invention provides a QA/QC method for filtering out inconsistencies across high-throughput bioassay processes, particularly across different biochips and different diluents or concentrations of diluents used in electrospray techniques.

[0007] The present invention uses the Knowledge Discovery Engine ("KDE") to identify hidden patterns across a wide variety of serum samples and biochips to generate a control model and agnostic to the underlying disease processes in question. Electrospray, MALDI-TOF (Matrix Assisted Laser Desorption/Ionization-Time of Flight) mass spectra, or SELDI-TOF (Surface Enhanced Laser Desorption/Ionization-Time of Flight) mass spectra can be analyzed in this manner, for example. Alternatively, the invention may use the KDE to identify hidden patterns across a variety of serum to diluent concentrations to generate a control model. In yet another embodiment, the KDE may be used to identify hidden patterns across a variety of diluents and sera samples to generate a control model.

[0008] The KDE is disclosed in U.S. Patent Application Serial No. 09/883,196, now U.S. Application Publication No. 2002/0046198A1, entitled "Heuristic Methods of Classification," filed June 19, 2001 ("Heuristic Methods"), and U.S. Patent Application Serial No. 09/906,661, now U.S. Application Publication No. 2003/0004402A1, entitled "A Process for Discriminating Between Biological States Based on Hidden Patterns from Biological Data," filed July 18, 2001 ("Hidden Patterns"); the contents of both applications are hereby incorporated by reference in their entirety. Software running the KDE is available from Correlogic Systems, Inc., under the name Proteome Quest ™.

[0009] After the KDE is used to generate a control model, a test serum may be compared to the control model to determine if the spectra produced by the high-throughput bioassay process and the serum are acceptable.

[0010] The KDE will search for hidden or subtle patterns of molecular expression that are, in and of themselves, "diagnostic." The level of the identified molecular products is termed *per se* diagnostic, because the level of the product is diagnostic without any further consideration of the level of any other molecular products in the sample.

[0011] In the data cluster analysis utilizing the KDE, the diagnostic significance of the level of any particular marker, *e.g.*, a protein or transcript, is a function of the levels of the other elements that are used to calculate a sample vector. Such products are referred to as "contextual diagnostic products." The KDE's learning algorithm discovers wholly new classification patterns without knowing any prior information about the identity or relationships of the data pattern, *i.e.*, without prior input that a specified diagnostic molecular product is indicative of a particular classification.

[0012] If the spectrum produced by the biochip and the serum map to the control model, then the data obtained from mass spectrometry of the serum and biochip may be used for further analysis. If the spectrum produced by the biochip and the serum fail to map to the control model, the data is deemed uncertified, and new data must be acquired. Alternatively, if a spectrum produced by a serum sample and a diluent map to the control model, then the spectrum obtained from an electrospray process may be used for further analysis. By using this method, inconsistencies across bioassay processes may be avoided, thereby improving the reliability of data obtained using the bioassay process. Other advantages may also be realized from the methods disclosed herein, as would be obvious to the ordinarily skilled artisan.

BRIEF DESCRIPTION OF THE DRAWINGS

[0013] FIG. 1 is a flow chart illustrating a method of obtaining a control model.

[0014] FIG. 2 depicts a table having various serum/biochip combinations that may be used to obtain the control model.

[0015] FIG. 3 illustrates a method of comparing the test serum to the control model.

[0016] FIG. 4 is a depiction of mapping of exemplary tolerances in three-dimensional space according to one aspect of the present invention.

[0017] FIG. 5 is a table illustrating results obtained from the KDE using two different types of biochips and 256 sera.

[0018] FIG. 6 is a flow chart of an alternative embodiment of the present invention for use with an electrospray process.

DETAILED DESCRIPTION

[0019] Generally, the invention includes a method of obtaining a control model for use in a bioinformatics system and a method for comparing a test sample against the model for the purpose of QA/QC.

[0020] A method of obtaining a control model according to one aspect of the present invention is illustrated in FIG 1. To ensure a highly articulate model, a variety of serum samples are selected at step 100. The selection should include selecting serum from as diverse a group of individuals as possible. The more diverse the selected sera, the more articulate the control model will be. For example, sera may be taken from healthy males, healthy females, males afflicted with various diseases, females afflicted with various diseases, persons of different ages, and persons of different races.

[0021] Once the diverse group of sera has been selected, a group of different biochips is selected at step 110. The diverse group of biochips may include an anionic chip, a cationic chip, and an immobilized metal chip. The selection of chips may include at least one anionic chip and at least one cationic chip. However, in order to generate a workable model at least two chips should be selected. For example, one model could be generated using three types of chips: WCX2 (cationic exchange), SAX2 (anionic exchange), and IMAC3 (immobilized metal) surface enhanced laser desorption/ionization ("SELDI") chips.

[0022] After the initial selection of sera (100) and the selection of biochips (110), the sera are applied to the chips in step 120. After each serum is applied to the surface of a chip, then each chip and sera combination is analyzed by mass spectrometry at step 130 to obtain a spectral output characterized by mass to charge (m/z) values. After one spectrum is produced, the process is repeated for a different biochip/serum combination. Each time a spectrum is obtained

for a particular biochip/serum combination, a determination is made at step 140 of whether all chips have been analyzed.

[0023] After all of the chips have been analyzed, a determination is made at step 150 of whether all sera have been analyzed by mass spectrometry in combination with each chip type. For example, assume three sera are selected at step 100, and two biochips are selected (one cationic exchange biochip ("Biochip A") and one anionic exchange biochip ("Biochip B")). After the first serum is analyzed by mass spectrometry on Biochip A, a determination is made at step 140 of whether all biochips have been used. Finding that the first serum has not been used with Biochip B, the process is repeated starting with step 120.

[0024] When both Biochip A and Biochip B have been analyzed with the first serum, a determination is made at step 150 of whether there are any more sera remaining. If any more sera remain, the process is repeated for each of the biochips. In this example, the process will be repeated for each of Biochip A and Biochip B, with the second and third sera respectively.

[0025] The data for each of the spectra may be stored, such as in the table illustrated in FIG. 2. The table includes data for "*i*" sera and "*j*" chips. Each cell in the table contains mass spectra (MS) data associated with each chip type and the various types of serum. For example, cell $MS_{j,i}$ corresponds to the spectral data from chip "*j*" and serum sample "*i*". After all of the data have been obtained, the stored mass spectrum data can be input into the KDE to discover hidden patterns in the spectral data.

[0026] Next, the process of developing a biochip model using the KDE will be described.

[0027] The data from each of the mass spectra are input into the KDE. The KDE then seeks to identify clusters of data (hidden patterns) in n-dimensional space, where n is the number of mass to charge values selected from the spectra for analysis, and each spectrum can be mapped into the n-dimensional space using the magnitude of each of the selected mass to charge values in the spectrum (the combination of a mass to charge value and the magnitude of the spectrum at that value being a vector). The KDE seeks clusters that contain as many of the spectra as possible and that distinguish each of the biochips from the others. Each cluster of data will define a centroid that will be associated with a particular biochip. In the event that a number of

possible groupings or combinations of clusters are identified by the KDE, the user will select the most optimal grouping to define the biochip model. The selection process could be automated rather than being directly performed by the user. In either case, the cluster with the highest population of vectors can be identified by either the user or the system and that cluster can be designated as the control model.

[0028] After the model has been obtained, test sera may be run against the model to determine if the sera/biochip combination is certified for further analysis. One method of QA/QC using the biochip model obtained in FIG. 1 is depicted in FIG. 3.

[0029] First, in step 300 multiple samples from a test serum is applied to a biochip. The test serum could be serum intended for a cancer screening, for example. Then in step 310 the test serum samples are analyzed by mass spectrometry. The spectra obtained in step 310 are then mapped to the biochip model in step 320.

[0030] Mapping the spectrum to the biochip model is performed in manner similar to the mapping of a serum sample to a training data set to diagnose a particular disease state as described in the Hidden Patterns application. Mapping a spectrum to the biochip model includes determining the spectrum value for each of the n mass to charge ratios in the biochip mode, plotting the associated vectors in the model's n-dimensional space, and comparing them with the location of the model's centroid. The centroid is defined as the center of the cluster determined to have the highest population of vectors from the model development.

[0031] After the spectrum of the test sample is mapped to the biochip model, it is determined in a step 330 if the spectrum maps within a predetermined hypervolume centered on the centroid in the model associated with the tested biochip.

[0032] If the spectrum maps within the predetermined hypervolume, the spectrum is deemed certified for further analysis. If the map of the spectrum falls outside the predetermined hypervolume, the spectrum is not deemed certified and the sample must be reanalyzed.

[0033] A system employing the method of the present invention should operate within predetermined tolerances. In determining whether a spectrum for a sample maps to the model for the biochip used with the sample, the point to which the vectors from each sample spectrum

maps in the model's n-dimensional space maps are compared to the location of the centroid for the cluster associated with the selected biochip. The spectrum is considered to map to the model if it lies within a predetermined hypervolume centered on that centroid. In this embodiment, the hypervolume defined with the centroid as its center will exclude approximately 95% of the total hypervolume of the n-dimensional space. The content of a polytope or other *n*-dimensional object is its generalized volume (i.e., its "hypervolume"). Just as a three-dimensional object has volume, surface area and generalized diameter, an *n*-dimensional object. has "measures" of order 1, 2, ..., *n*, the hypervolume is defined based on these measures of order. The hypervolume can also be defined in terms of a predefined acceptable process tolerance. The n-dimensional hypervolume calculation is akin to Mahalanobis distances used in establishing rejection and acceptance criteria.

[0034] This can be visualized in three dimensions as depicted in FIG. 4. FIG. 4 illustrates a centroid "C," which is associated with the cluster of features plotted in n-dimensional space (here the space is depicted as three-dimensional for visualization purposes). A theoretical sphere, "S" is located in the n-dimensional space. The sphere is centered at the location of the centroid "C." Tolerances should be set such that the sphere "S" has a volume that excludes approximately 95% of the volume of the three-dimensional space. The volume of the three dimensional space is defined by the set of plotted features in that space, and is bounded by (and preferably normalized to) the m/z feature with the greatest intensity.

[0035] Referring back to FIG. 3, once a sample is deemed certified, the spectral data may be used for further analysis. The further analysis may entail running the data through the KDE to discover hidden patterns in the spectra. These hidden patterns may be compared to disease state models to diagnose for a particular disease. This method of diagnosis is described in further detail in the Hidden Patterns application.

[0036] If no spectra map to the model, then in step 350 the sample and biochip are deemed non-certified. If the sample and the chip were deemed non-certified the for the first time using a particular serum sample, then in steps 360 and 370 a determination is made to see if the "no certification" is a first "no certification"; is so, a new chip is selected and the process repeated.

[0037] If the sample and the biochip were determined to be non-certified more than once, then in step 380 a new serum sample and chip are obtained, and the process is repeated. After a certified sample has been obtained, at step 410 it is output for further analysis.

[0038] In one exemplary test, sera was obtained from subjects with and without cancer. Two types of biochips were selected, IMAC3 (immobilized metal) and WCX2 (cationic exchange). A diverse group of 256 sera were selected. The sera were then applied to each type of chip, analyzed by mass spectrometry, and the spectral output collected. The spectral data was input into the KDE.

[0039] The model identified by the KDE is shown in FIG. 5. The table in FIG. 5 shows the constituent "patterns" or clusters comprising the model. Each cluster corresponds to a point, or node, in the N-dimensional space defined by the N *m/z* values (or "features") included in the model. In this case, 10 m/z values are included in the model, so n=10. The table shows the constituent centroids of the mode, each in a row identified by a "Node" number. Thus, this model has eight nodes or centroids. The table also includes columns for the constituent features or vectors of the centroids, with the *m/z* value for each vector identified at the top of the column. The amplitudes are shown for each feature or m/z value, for each centroid, and are normalized to 1.0.

[0040] The remaining four columns in the table are labeled "Count," "State," "StateSum," and "Error." "Count" is the number of samples in the Training set that correspond to the identified node. "State" indicates the state of the node, where 1 indicates the IMAC3 chip and 0 indicates the WCX2 chip. "StateSum" is the sum of the state values for all of the correctly classified members of the indicated node, while "Error" is the number of incorrectly classified members of the indicated node. Thus, for node 2 , 108 samples were assigned to the node, whereas 104 samples were actually from the IMAC3 chip. StateSum is thus 104 (rather than 108) and Error is 4.

[0041] The cluster that contained the highest population of vectors was designated as the control model. In the Example illustrated in FIG. 5, node 2 defined the control model because it contained the greatest number of vectors (108).

[0042] While the method described above has been described for use with biochips, another embodiment of the invention may be practiced using electrospray techniques to obtain data relating to a particular serum.

[0043] When using an electrospray technique for obtaining biological data for diagnostic purposes, a primary factor limiting the consistency of the data obtained is the particular diluents used in preparing the sample for electrospray analysis. Therefore, rather than characterizing the serum in conjunction with the biochips being used, electrospray techniques used in combination with the present methods will characterize the serum in combination with the diluents used.

[0044] To use electrospray ionization ("ESI") to obtain accurate spectral data, a stable spray should be obtained. There are three physical characteristics of the spectral data that yield spectral results that can be utilized by the KDE. These physical characteristics include: (1) the number of mass peaks; (2) the total ion current; and (3) the stability of the spray.

[0045] Various tests were run to determine a preferred diluent concentration and composition to yield effective results for use in the KDE. These tests were run using electrospray apparatus manufactured by Advion BioSciences, and particularly the NanoMate100™ ESI in conjunction with Correlogic's Proteome Pattern Blood Test™ which is based on the Proteome Quest™ software.

[0046] One particular test involved diluting the serum sample at 1:1000 in 50:50 acetonitrile: H_2O containing 0.2% formic acid (FA). While 20 λ of each sample was aliquoted into each well, representing 0.02 λ of serum, only about 100-200 nano-liters was actually sprayed and analyzed. Operating the Q Star mass spectrometer in positive ion TOF/MS mode, data was acquired for 2 minutes/sample, m/z of 300-2000, using a 1 second scan rate. The nanospray was initiated by applying 1.55 kV spray voltage at a pressure of about 0.5 psi.

[0047] Based on testing of serum samples a determination was made that the Multichannel acquisition mode (MCA) should be used on the Q-star mass spectrometer when obtaining spectral data. This is based on a determination that the MCA mode produced better resolution of the spectral peaks.

[0048] To optimize the diluent used in preparing the serum samples, various other tests were run. The concentrations of the serum to diluent are preferably between 1:1000 and 1:250, but other diluent concentrations may be selected in a manner apparent to those skilled in the art. Two diluent types tested included acetonitrile (ACN) and methanol (MeOH). Each of these diluents was combined with an acid. Acids include, but are not limited to, trifluoroacetic acid (TFA), formic acid (FA), and acetic acid. Either TFA or FA can be used for purposes of the present invention and are preferably in concentrations between about 0.2 % acid and 1.0 % acid.

[0049] An alternative embodiment of the method of obtaining a model according to the present invention will now be described in relation to FIG. 6. The step of selecting sera in step 100 is the same as described above. Again, the more diverse the overall group of selected sera is, the more articulate the model will be.

[0050] Secondly, in step 610 a selection of diluents is made. Diluents selected may be diverse or homogeneous. For example, a diverse group of diluents including ACN and MeOH may be selected. Alternatively, only ACN may be selected as long as the concentrations of the ACN differ (e.g., 1:1000, 1:500, and 1:250 serum to ACN).

[0051] Thirdly, in step 620, a mixture is made according to predetermined concentrations of serum to diluent. This mixture is then analyzed using electrospray in step 630 to obtain spectral data. This process is repeated for all desired concentrations, diluents and serum samples until all data have been obtained in steps 640, 650, and 660. A model is then obtained in step 670 based on the data extracted from the various samples. The model is obtained using the KDE in the same manner as described above for the acquisition of a biochip model.

[0052] A method of QA/QC using the electrospray is substantially the same as for the disclosed method of QA/QC for biochips. Notable variations can be found in the generation of the sample and the method of obtaining the data. The significant differences between the overall processes stems from the differences in obtaining the model and its ability to identify a particular serum diluent.

[0053] In one particular test, sera were obtained from male and female subjects. Two diluents, acetonitrile (ACN) and methanol (MeOH) were selected. A mixture was made at a

concentration of 1:250 of serum to diluent. Selected sera mixture (102 samples) were analyzed by electrospray mass spectrometry, and the spectral output collected. The spectral data was input into the KDE. The KDE identified a model containing three clusters in total that distinguished the two diluents. One cluster was associated with ACN and that cluster was designated as the control model.

[0054] The above-described method is applicable to various bioassay processes to ensure that both the particular high-throughput bioassay process being used and the serum being tested will yield an accurate diagnostic. By using the method described above, biological diagnostics may be provided that have increased accuracy and reliability.

[0055] Tolerances to employ the aforementioned methods were described as being such that a hypervolume defined about the centroid of a cluster that will exclude approximately 95% of the total hypervolume of the n-dimensional space. While 95% percent was explicitly mentioned, one of ordinary skill in the art would realize that the methods of the present invention would operate effectively with different sized hypervolumes centered on the centroid.

[0056] In the described embodiments, biochips and electrospray processes were illustrative. Various other high-throughput bioassay processes are known in the art and could be employed with the methods of the present invention.

[0057] In the described embodiments, to obtain a model characteristic of a particular high-throughput bioassay process, sera should be taken from healthy males, healthy females, males afflicted with a disease, females afflicted with a disease, persons of different ages and persons of different races. While these specific examples were given, numerous other diverse sera samples could be taken. The best possible diverse sera would contain serum from every individual in the world. Therefore, taking sera from any individual that does not group into one of the aforementioned classifications is within the scope of the present invention.

[0058] While specific diluents and acids were described in reference to the methods of QA/QC for electrospray techniques, these diluents and acids are not intended to be exhaustive and a variety of other suitable diluents and acids are suitable for those explicitly mentioned. Additionally, while specific concentrations of both acids and ratios of sera to diluent were

disclosed, one of ordinary skill in the art will realize the specific concentrations will depend on the particular acids and diluents used to perform the inventive method, and the described acids and diluents are not intended to be all inclusive. Various other concentrations in combination with various acids and diluents will be obvious to the ordinarily skilled artisan based on the teachings of the present invention.

[0059] The various features of the invention have been described in relation to a method of quality assurance/quality control of high-throughput bioassay processes. However, it will be appreciated that many of the steps may be implemented with various apparatus and bioinformatics methods. Moreover, variations and modifications exist that would not depart from the scope of the invention.